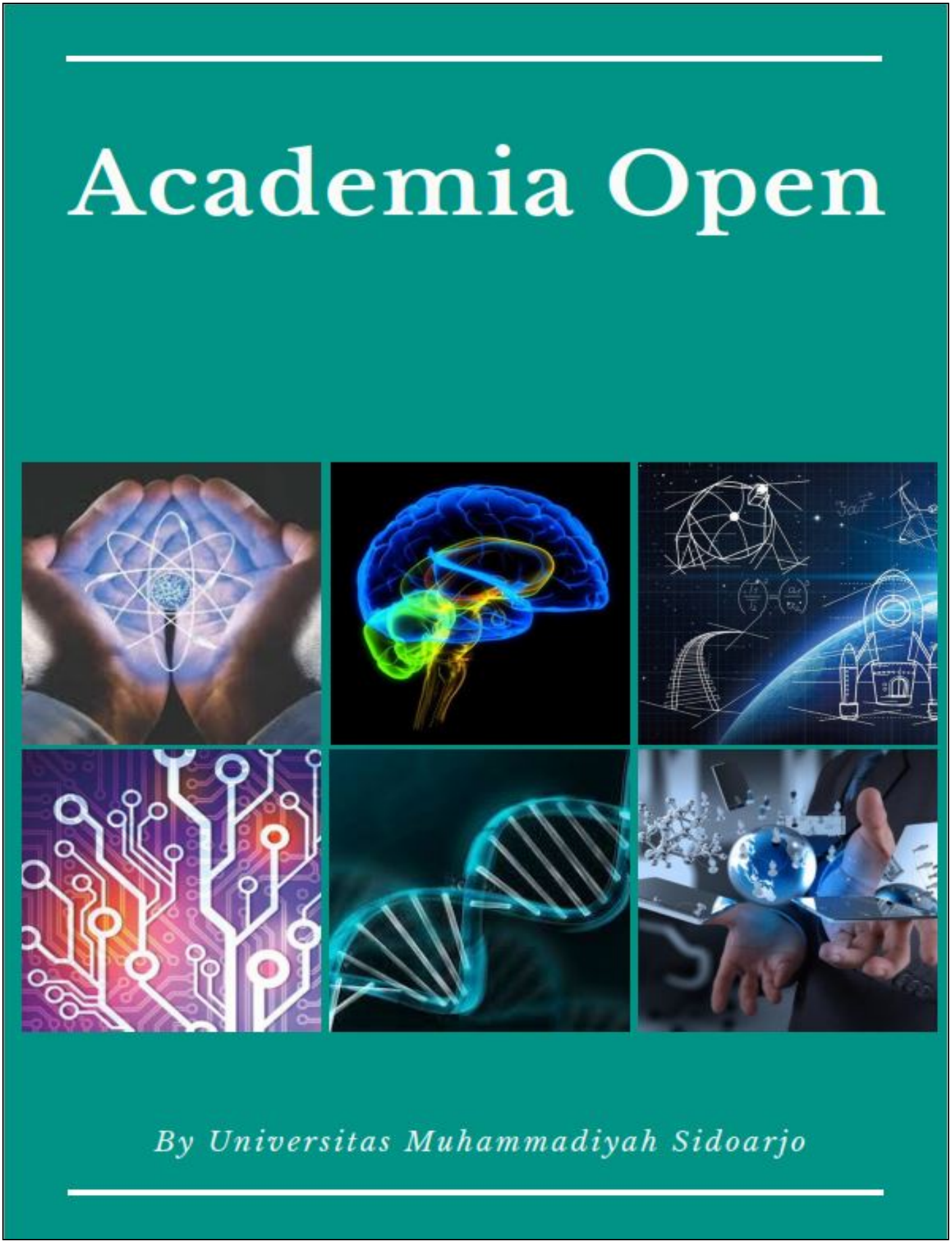
Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

Table Of Content

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)



ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

Originality Statement

The author[s] declare that this article is their own work and to the best of their knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the published of any other published materials, except where due acknowledgement is made in the article. Any contribution made to the research by others, with whom author[s] have work, is explicitly acknowledged in the article.

Conflict of Interest Statement

The author[s] declare that this article was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright Statement

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

EDITORIAL TEAM

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

Article information

Check this article update (crossmark)

Check this article impact (*)

Save this article to Mendeley

(*) Time for indexing process is various, depends on indexing database platform

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

Sentiment Analysis of Potential Presidential Candidates 2024: A Twitter-Based Study

Analisis Sentimen Calon Presiden Potensial 2024: Sebuah Studi Berbasis Twitter

Yulian Findawati, yulianfindawati@umsida.ac.id, (1)
Universitas Muhammadiyah Sidoarjo, Indonesia

Uce Indahyanti, yulianfindawati@umsida.ac.id, (0)
Universitas Muhammadiyah Sidoarjo, Indonesia

Yunianita Rahmawati, yulianfindawati@umsida.ac.id, (0)
Universitas Muhammadiyah Sidoarjo, Indonesia

Ratih Puspitasari, yulianfindawati@umsida.ac.id, (0)
Universitas Muhammadiyah Sidoarjo, Indonesia

(1) Corresponding author

Abstract

This study aims to analyze the sentiment towards potential presidential candidates for the 2024 election in Indonesia based on Twitter users' opinions. Three prominent figures, Ganjar Pranowo, Anies Baswedan, and Prabowo Subianto, were surveyed to gauge their electability. Using machine learning classification methods, Support Vector Machine, Bernoulli Naïve Bayes, and Logistic Regression, sentiment classification was performed. The findings indicate that Twitter users expressed predominantly positive sentiments towards each potential candidate. The evaluation of the classification algorithms showed SVM with 84% accuracy, Bernoulli Naïve Bayes with 77%, and Logistic Regression with 84%. This research sheds light on public sentiment towards potential leaders, offering valuable insights for political strategists and decision-makers in shaping effective election campaigns.

Highlight:

- Sentiment Analysis: The study employs machine learning techniques to analyze the sentiments expressed by Twitter users towards potential presidential candidates for the 2024 election in Indonesia.
- Positive Sentiments: The findings reveal that Twitter users predominantly exhibit positive sentiments towards all three potential candidates, Ganjar Pranowo, Anies Baswedan, and Prabowo Subianto.
- Election Insights: This research provides valuable insights into public sentiment,

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

offering valuable information for political strategists and decision-makers in devising effective election campaigns for the upcoming presidential election.

Keyword: Sentiment Analysis, Twitter Users, Potential Presidential Candidates, Machine Learning, Election 2024

Published date: 2023-08-04 00:00:00

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

Pendahuluan

Indonesia merupakan negara yang menganut sistem demokrasi. Hal ini ditandai dengan sistem pemilihan umum dengan menganut suara terbanyak baik terhadap pemilihan umum Presiden, Wakil Rakyat, Gubernur, Bupati, maupun Kepala Desa. Pemilihan umum pada suatu negara yang menganut asas demokrasi biasanya diselenggarakan secara periodik. Bagi seluruh rakyat Indonesia, tahun 2024 akan menjadi tahun pesta demokrasi terbesar, karena pada tahun tersebut akan menandai berakhirnya masa jabatan Presiden dan Wakil Presiden Indonesia saat ini. Banyak sekali autisiasme masyarakat Indonesia terkait pasangan yang baik di calonkan ataupun mencalonkan diri baik di dunia nyata maupun di dunia maya melalui media sosial seperti Twitter. Twitter dapat dianggap sebagai media bagi pengguna dan kandidat untuk mendapatkan eksposur yang luas sekaligus mengekspresikan pendapat mereka kepada khalayak global. Untuk menentukan kandidat mana yang paling cocok untuk memimpin Indonesia selama lima tahun ke depan. Perkembangan massif jejaringan social saat ini sering terjadi sebagai dasar untuk salah satu alat kampanye dalam melakukan kegiatan politisasi, sehingga kehadiran media sosial ini dapat digunakan oleh berbagai kalangan masyarakat Indonesia dalam beberapa tahun terakhir. Oleh sebab itu, penting untuk menganalisis dan memahami peran yang dapat dimainkan Twitter dalam mengukur sentimen seputar topik-topik penting[1].

Menurut sebuah laporan yang diterbitkan oleh Wearsocial pada awal 2018 menjelaskan bahwa pengguna internet di Indonesia mencapai 132 juta dengan persentase 60%. Akses internet melalui smartphone menjadikan Indonesia urutan keempat negara dengan akses internet, untuk penggunaan media sosial, Indonesia berada di peringkat ketiga dengan 53 juta pengguna. Menurut data yang dirilis oleh Twitter Indonesia pada akhir tahun 2016, sebesar 77% pengguna Twitter di Indonesia dikatakan sebagai pengguna aktif. Terdapat komentar positif, negatif maupun netral dari mayarakat menjelang pilpres maupun ketika pilpres sedang berlangsung yang diselenggarakan mendatang. Ulasan twitter masih belum terindentifikasi adalah review positif, negatif maupun netral sebuah informasi yang diterima langsung dari media Twitter. Pada Twitter bisa didapatkan opini, keinginan atau komentar dari sosial dapat digunakan untuk mengungkapkan peristiwa yang sedang terjadi dalam kasus ini terkait bakal calon presiden dengan #GanjarPranowo #AniesBaswedan dan #PrabowoSubianto. Agar pendapat ini dapat digunakan dan bermanfaat, diperlukan keragaman. Proses untuk mendapatkan informasi penting melalui analisis sentimen[2].

Analisis sentimen adalah Teknik untuk menganalisis opini, sentimen, penilaian dan emosi untuk entitas seperti produk, layanan, acara atau atribut lainnya. Pemikiran dasar dari teknik analisis sentimen adalah untuk lebih mengklasifikasi teks, kalimat atau dokumen mengidentifikasi teks, frasa atau materi yang terkandung dalam sentimen atau opini positif, negatif atau netral[3]. Dalam penelitian ini, analisis sentimen dilakukan dengan melihat dan mengumpulkan informasi terkait opini serta pendapat masyarakat Indonesia dalam berbahasa Indonesia dengan menggunakan bantuan sosial media Twitter yang dimaksudkan kepada Ganjar Pranowo, Anies Baswedan dan Prabowo Subianto menjadi kandidat dalam pemilihan presiden untuk tahun 2024, apakah opini yang beredar didalam Twitter tergolong kategori opini positif, netral atau bahkan negatif. Dalam artikel ini, akan dibahas analisis mendalam tentang data yang diambil dari media sosial Twitter terkait calon presiden yang sering disebut-sebut. Artikel ini akan menyajikan dan menampilkan data serta membahas bagaimana persepsi public tehadap capres Indonesia 2024. Dari hasil analisis ini diharapkan dapat mengetahui trend penilaian capres dari komunitas Twitter. Dari analisis ini, dapat diharapkan bisa sebagai tambahan referensi politik yang nantinya akan diselenggarakan pada tahun 2024. Mengingat pilpres 2024 merupakan agenda besar yang menarik untuk dibahas[4].

Dengan demikian, diperoleh hasil dari model tersebut machine learning memiliki kemampuan yang berbeda untuk disalahartikan karena label yang dihasilkan tidak bisa dianggap sebagai “kebenaran mendasar”. Label dalam konteks ini bahkan tidak akan menjawab hipotesis ini karena ada tingkat subjektivitas yang terlibat dalam menjelaskan sentimen yang ada dalam sosial media Twitter. Menentukan label positif, negatif maupun netral dapat dilakukan secara otomatis menggunakan analisis sentimen VADER (*Valence Aware Dictionary and Sentiment Reasoner*) yang merupakan model untuk analisis sentimen teks yang peka terhadap polaritas (positif/negatif/netral) dan intensitas (kekuatan) sentimen. Analisis sentimen akan menentukan apakah opini yang diungkapkan dalam kalimat atau dokumen itu positif, negatif atau netral dengan klasifikasi polarisasi teks. Sehingga jika pelabelan manual yang biasa digunakan dalam analisis sentimen dinilai kurang efisien dari segi waktu dan tenaga, terutama jika data yang digunakan dalam jumlah yang besar[5] .

Dengan analisis sentimen metode yang diperlukan untuk mendukung klasifikasi. Metode yang digunakan adalah Support Vector Machine (SVM), berdasarkan hasil penelitian analisis sentimen sebelumnya yang dilakukan oleh Wanda, Imam dan Rizal. Pada penelitian tersebut dilakukan analisis sentimen pada objek Twitter dengan mengimplementasikan metode SVM dari pengujian tersebut didapatkan hasil akurasi tertinggi sebesar 90% pada komposisi data train 50% dan komposisi data uji 50%[6]. Penelitian lain tentang sentimen analisis yang pernah dilakukan pada bulan April 2023 menggunakan metode Bernouli Naïve Bayes, Support Vector Machine, dan Logistic Regression. Dengan hasil akurasi dari metode Bernouli Naïve Bayes 76% sedangkan metode Support Vector Machine menghasilkan akurasi 92%. Adapun metode Logistic Regression menghasilkan akurasi 92%[7]. Penelitian serupa juga dilakukan oleh yang meneliti tentang ibu kota akan dipindahkan ke Kalimatan Timur pada 26 Agustus 2019. Perencanaan tesebut menuai pro dan kontra dari masyarakat. Metode klasifikasi yang digunakan untuk melakukan sentimen analisis adalah Naïve Bayes Classifier untuk model Bernoulli dan Multinomial. Hasil menunjukkan bahwa Bernoulli Naïve Bayes memiliki tingkat sensitivitas (recall) 93,45% dan Multinomial Naïve

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Bayes memiliki tingkat sensitivitas (recall) 90,19% artinya baik Bernoulli maupun Multinomial memiliki hasil yang baik untuk penelitian tersebut[8].

Pada penelitian ini metode pembelajaran mesin yang digunakan untuk mengklasifikasikan opini dari data yang sangat banyak tersebut. Untuk melakukan klasifikasi data tersebut peneliti menggunakan metode *Support Vector Machine, Bernoulli Naïve Bayes* dan *Logistic Regression*.Penelitian ini bertujuan untuk mengevakuasi model klasifikasi sentimen dari kombinasi algoritma *Support Vector Machine, Bernoulli Naïve Bayes, dan Logistic Regression*.Komparasi ketiganya menjadi menarik untuk diteliti karena digunakan sebagai pembanding saja. Selain itu kajian penelitian tentang metode-metode analisis sentimen diperlukan khususnya mengevaluasi akurasi dari konfigurasi ketiga model klasifikasinya.

Metode

Pada penelitian ini terdapat langkah-langkah dalam penelitian yang dilakukan sebagaimana digambarkan seperti gambar 1.

A. *Crawling*

Tahap crawling data pada penelitian ini memiliki tujuan untuk mengumpulkan atau mengunduh data dari database. Dengan memanfaatkan Netlytic. Netlytic merupakan web pengumpulan data dan penganalisa teks yang dapat meringkat data teks secara otomatis dan menemukan jaringan komunikasi dari postingan media sosial yang tersedia untuk umum dan juga merupakan salah satu aplikasi yang membantu penelitian jaringan komunikasi. Kelebihan Netlytic diantaranya mampu menangkap komentar di Twitter, Youtube, RSS Feed atau file teks/CSV. Prosedur penggunaan Netlytic dimulai dari membuat akun kemudian mengimport data yang dianalisis. Akun dengan level kedua sudah cukup sebagai tempat untuk meneliti komentar dengan jumlah kurang dari 10.000. Tidak memiliki perbedaan ketepatan analisis antara akun tingkat satu, kedua, dan ketiga. Perbedaannya hanya pada jumlah akun tingkat ketiga yang dapat menguji lebih banyak data yaitu 10.000. Data yang diperoleh berupa opini dituliskan dalam Bahasa Indonesia, yaitu tweet dengan kata kunci “#GanjarPranowo #AniesBaswedan dan #PrabowoSubianto” yang sebagai calon bakal presiden 2024. Pengambilan data dimulai dari tanggal 7 Mei 2023-14 Mei 2023 [9].

B . Preprocesing Data

Sebelum dapat mengklasifikasikan data yang diperoleh melalui proses crawling, diperlukan langkah pre-processing agar data tersebut menjadi lebih terstruktur dan bersih. Hal ini dikarenakan data awal sering kali mengandung banyak symbol dan kata-kata yang tidak relevan. Dengan melakukan pre-processing data, dalam penelitian ini dapat memastikan bahwa data menjadi terorganisir sehingga memungkinkan untuk dilakukan proses klasifikasi[10]. Pada penelitian ini diperlukan tahap ini agar opini atau ulasan yang didapatkan dari data tweet hanya pendapat mengenai calon bakal presiden 2024 saja. Selanjutnya dilakukan proses cleansing data melibatkan serangkaian tahapan yang kompleks, termasuk menetapkan aturan untuk mengidentifikasi kualitas data, mengolah kesalahan atau kecacatan data, serta melakukan perbaikan atas kesalahan tersebut. Setelah itu dilakukan tahap stemming yaitu salah satu yang perlu dilakukan untuk meminimalkan jumlah indeks data yang berbeda sehingga dengan akhiran atau awalan kembali ke bentuk dasarnya[11].

C. Tokenisasi

Tahap tokenisasi atau biasa disebut parsing adalah tahapan yang diimplementasikan setelah data dikumpulkan pada tahap cleaning data. Pada tahap ini, dilakukan penghapusan tanda baca dan pemisahan kata-kata berdasarkan indeks spasi untuk memperoleh hasil yang lebih terstruktur [12].

D . Load Distionary

Setelah melakukan tahap tokenisasi terhadap data, langkah berikutnya adalah mengklasifikasikan kata-kata ke dalam kelompok yang sesuai dengan makna asli dari kata yang dimaksud. Tahap Load Dictionary merupakan tahapan dimana kamus dengan kata kunci yang mengindikasikan sentimen positif, negatif, atau netral[13].

E . Visualisasi

Setelah berhasil menyelesaikan keempat tahapan sebelumnya, tahap terakhir adalah melakukan visualisasi data terkait sentimen yang ditulis oleh akun pribadi atau media social tentang calon bakal pemilihan presiden 2024. Visualisasi data dapat dilakukan dengan menggunakan WordCloud dan Bar Chart[14]. Tujuan dari visualisasi untuk memperoleh informasi spesifik dari sekumpulan data yang ada. Analisis sentimen difokuskan pada pengolahan ulasan atau pendapat yang memiliki skor polaritas, yakni nilai sentimen positif, negatif, atau netral. Sehingga dapat diambil kesimpulan dari hasil yang didapatkan [15].

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

Hasil dan Pembahasan

A. Twitter Crawling

Dataset yang digunakan dalam penelitian ini berupa format .csv diperoleh dari database Twitter dengan menggunakan teknik scrapping menggunakan Netlytic. Data yang dihasilkan adalah data acak terlepas dari pemilik akun baik asli maupun dimiliki oleh media dan akun yang terindikasi palsu. Opini masyarakat terkait calon bakal presiden 2024 yang sudah berhasil terkumpulkan sebanyak 2500 masing-masing tiga kata kunci yaitu "#GanjarPranowo #AniesBaswedan dan #PrabowoSubianto".

```
# Menampilkan data sebanyak 500

tweet_df
```

link	author	title	description	pubdate	source	favorite_count	...	user_id
r.com/mawardi_affan/statuses/165...	mawardi_affan	RT @idtodayco: BERANI JUJUR ITU HEBAT! Siapa ...	BERANI JUJUR ITU HEBAT! Siapa menurut anda Ca...	2023-05-07 10:37:30	Twitter for Android	0	...	1401560067011317768
r.com/TrioKw3kkw3k__/statuses/16...	TrioKw3kkw3k__	RT @P3n99u94t: SAFARI POLITIK DINILAI TAK ETIS...	SAFARI POLITIK DINILAI TAK ETIS, GANJAR BAIKNY...	2023-05-07 10:37:30	Twitter for Android	0	...	1647301174503297025
r.com/adis01saputra/statuses/165...	adis01saputra	@yoyokdenbagoes @Lacak08 @AndreasSolusi @are_j...	@yoyokdenbagoes @Lacak08 @AndreasSolusi @are_j...	2023-05-07 10:37:24	Twitter Web App	0	...	1638462815777476609
com/dam_yanto/statuses/1655220...	dam_yanto	RT @idtodayco: Coba Polling Kotak Kosong sama ...	Coba Polling Kotak Kosong sama Pak GP. Seanda...	2023-05-07 10:37:20	Twitter Web App	0	...	1449033983874924554
com/PecelWonosobo/statuses/165...	PecelWonosobo	RT @Subur0204: Saat saya buat polling dan Anis...	Saat saya buat polling dan Anis menang, mereka...	2023-05-07 10:37:19	Twitter for Android	0	...	1577231438990057472

Figure 1. *Data mentah hasil twitter crawling*

Data hasil twitter crawling kemudian akan difilter menjadi satu kolom saja. Kolom yang akan digunakan hanya kolom description yaitu kolom ulasan untuk mempermudah analisis proses ketahap berikutnya.

B . *Pre-Processing Data*

Pre-processing data adalah proses mempersiapkan data untuk dimasukkan ke dalam model machine learning. Hasil pengumpulan data biasanya masih banyak noise dan sulit dimengerti, sehingga data perlu dibersihkan dan mentransformasi data agar siap untuk diolah. Berikut merupakan langkah-langkah dalam melakukan pre-processing data.

1. *Case Folding* adalah proses mengubah data tweet menjadi huruf kecil (*lowercase*). Berikut merupakan contoh data penelitian yang dilakukan pada proses *case folding*.

2.

```
df['Remove_RT'] = df['Remove_RT'].str.lower()
df
```

Figure 2. *Source code case folding*

Tweet	Hasil Case Folding
BERANI JUJUR ITU HEBAT! Siapa menurut anda Capres yang mampu melaksanakan "Keadilan Sosial Bagi Seluruh Rakyat Indonesia"? Apakah Ganjar atas Anies yg berpotensi mampu menjawab tantangan politik Global? Ganjar = Like Anies = Retweet https://t.co/VCbzJxcc5H	berani jujur itu hebat siapa menurut anda capres yang mampu melaksanakan keadilan sosial bagi seluruh rakyat indonesia apakah ganjar atas anies yg berpotensi mampu menjawab tantangan politik global ganjar like anies retweet
SAFARI POLITIK DINILAI TAK ETIS, GANJAR BAIKNYA PERTIMBANGKAN MUNDUR DARI GUBERNUR...!! #FirliUsutKorupsiGanjar #FirliUsutKorupsiGanjar □□□□ https://t.co/laQljVagII	safari politik dinilai tak etis ganjar baiknya pertimbangkan mundur dari gubernur
Pak Ganjar Bisa Sesuaikan Cara Kerja Jokowi, Kalo Pak	pak ganjar bisa sesuaikan cara kerja jokowi kalo pak

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

Prabowo Sorry Nih Agak Susah Karena Umur. □□	prabowo sorry nih agak susah karena umur

Table 1. *Sampel tweet hasil case folding*

3. *Cleaning* merupakan langkah untuk menghapus karakter yang tidak perlu seperti URL, @, #, https:, RT (Retweet), angka, symbol dan emoticon.

```
#-----------Cleaning Data-----------------

def remove(tweet):
    #remove angka
    tweet = re.sub(r"\d+", "", tweet)
    # tweet = re.sub('[0-9]+', '', tweet)

    # remove stock market tickers like $GE
    tweet = re.sub(r'\$\w*', '', tweet)

    # remove old style retweet text "RT"
    tweet = re.sub(r'^RT[\s]+', '', tweet)

    # remove hashtags
    # only removing the hash # sign from the word
    tweet = re.sub(r'#([A-Za-z0-9_]+)', '', tweet)

    # remove hyperlinks
    tweet = re.sub(r'https?:\/\/.*[\r\n]*', '', tweet)

    #remove coma
    tweet = re.sub(r',','',tweet)

    # remove non ASCII (emoticon, chinese word, .etc)
    tweet = tweet.encode('ascii', 'replace').decode('ascii')

    # remove tab, new line, ans back slice
    tweet = tweet.replace('\\t'," ").replace('\\n'," ").replace('\\u'," ").replace('\\',"")

    # remove hashtags
    tweet = re.sub(r'#', '', tweet)
```

Figure 3. *Source code cleaning*

Tweet	Hasil Cleaning
BERANI JUJUR ITU HEBAT! Siapa menurut anda Capres yang mampu melaksanakan "Keadilan Sosial Bagi Seluruh Rakyat Indonesia"? Apakah Ganjar atas Anies yg berpotensi mampu menjawab tantangan politik Global? Ganjar = Like Anies = Retweet https://t.co/VCbzJxcc5H	berani jujur itu hebat siapa menurut anda capres yang mampu melaksanakan keadilan sosial bagi seluruh rakyat indonesia apakah ganjar atas anies yg berpotensi mampu menjawab tantangan politik global ganjar like anies
SAFARI POLITIK DINILAI TAK ETIS, GANJAR BAIKNYA PERTIMBANGKAN MUNDUR DARI GUBERNUR...!! #FirliUsutKorupsiGanjar #FirliUsutKorupsiGanjar □□□□ https://t.co/laQljVagII	safari politik dinilai tak etis ganjar baiknya pertimbangkan mundur dari gubernur
Pak Ganjar Bisa Sesuaikan Cara Kerja Jokowi, Kalo Pak Prabowo Sorry Nih Agak Susah Karena Umur. □□	pak ganjar bisa sesuaikan cara kerja jokowi kalo pak prabowo sorry nih agak susah karena umur

Table 2. *Sampel tweet hasil cleaning*

Proses setelah dilakukan cleaning data yaitu menghapus duplikat kata. Proses duplikat kata dilakukan untuk menganalisis kata-kata yang penting. Setelah menyelesaikan proses drop duplikat jumlah data tweet menjadi 975 pada dataset pak Ganjar, sedangkan dataset pak Anies menjadi 367, dan dataset pak Prabowo menjadi 157.

4. *Tokenizing* dalam penelitian ini merupakan langkah dalam memecah string atau input terhadap teks yang telah melalui langkah cleaning berdasarkan tiap kata yang menyusunnya dan menghilangkan URL, @mention dan hashtag. Tahap tokenization dilakukan dengan menggunakan fungsi nltk_tokenize(), sebuah library dalam bahasa pemrograman Python3 bernama NLTK (Natural Language Tool Kit). NLTK merupakan sebuah library yang digunakan untuk memproses teks seperti melakukan *classification, tokenization, stemming, tagging, parsing dan semantic reasoning*.

```
# Mendefinisikan fungsi untuk tokenization

#-------Tokenizing----------
from nltk.tokenize import TweetTokenizer

import re
def tokenization(tweet):
    tokens = re.split(r"\s+",tweet,)
    return tokens

# Menerapkan fungsi ke kolom
df['Tweet_Tokenizer']= df['Remove_RT'].apply(lambda x: tokenization(x))
df
```

Figure 4. *Source code tokenizing*

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

Tweet	Hasil Tokenizing
berani jujur itu hebat siapa menurut anda capres yang mampu melaksanakan keadilan sosial bagi seluruh rakyat indonesia apakah ganjar atas anies yg berpotensi mampu menjawab tantangan politik global ganjar like anies	[berani,jujur,itu,hebat,siapa,menurut,anda,capres,yang,mampu,melaksanakan,keadilan,sosial,bagi,seluruh,rakyat,indonesia,apakah,ganjar,atas,anies,yg,berpotensi,mampu,menjawab,tantangan,politik,global,ganjar,like,anies]
safari politik dinilai tak etis ganjar baiknya pertimbangkan mundur dari gubernur	[safari,politik,dinilai,tak,etis,ganjar,baiknya,pertimbangkan,mundur,dari,gubernur]
pak ganjar bisa sesuaikan cara kerja jokowi kalo pak prabowo sorry nih agak susah karena umur	[pak,ganjar,bisa,sesuaikan,cara,kerja,jokowi,kalo,pak,prabowo,sorry,nih,agak,susah,karena,umur]

Table 3. *Sampel tweet hasil tokenizing*

5. *Stopword removal* merupakan langkah untuk menghilangkan kata informatif rendah. *Stopword* dilakukan jika kalimat berisi kata-kata umum dan tidak penting seperti waktu, penghubung, dan sebagainya.

```
#-------------Stop word removal----------------
from nltk.corpus import stopwords
list_stopwords =  stopwords.words('indonesian')

list_stopwords.extend(['dengan', 'ia', 'bahwa', 'oleh', 'pas', 'ohiya', 'y', 'a', 'deh', 'loh', 'ya',
                       'sih', 'oh', 'ah', 'ok', 'nih', 'gih', 'kan'])


data = pd.read_csv('stopwords-indonesia.csv', names=["stopwords"], header= None)
list_stopwords.extend(data["stopwords"][0].split(' '))
set(list_stopwords)
def remove_stopwords(words):
    return [word for word in words if word not in list_stopwords]

# Menerapkan hasil stopwords di penambahan kolom yang bernama (stopwords)
df['Stopwords']= df['Tweet_Tokenizer'].apply(lambda x:remove_stopwords(x))
df
```

Figure 5. *Source code stopword removal*

Tweet	Hasil Stopword Removal
berani jujur itu hebat siapa menurut anda capres yang mampu melaksanakan keadilan sosial bagi seluruh rakyat indonesia apakah ganjar atas anies yg berpotensi mampu menjawab tantangan politik global ganjar like anies	[berani,jujur,hebat,capres,melaksanakan,keadilan,sosial,rakyat,indonesia,ganjar,anies,yg,berpotensi,tantangan,politik,global,ganjar,like,anies]
safari politik dinilai tak etis ganjar baiknya pertimbangkan mundur dari gubernur	[safari,politik,dinilai,etis,ganjar,baiknya,pertimbangkan,mundur,gubernur]
pak ganjar bisa sesuaikan cara kerja jokowi kalo pak prabowo sorry nih agak susah karena umur	[ganjar,sesuaikan,kerja,Jokowi,kalo,Prabowo,sorry,susah,umur]

Table 4. *Sampel tweet hasil stopword removal*

Pre-processing data dilakukan untuk mengubah tweet atau data teks yang tidak terstruktur sehingga data tersebut dapat disusun sesuai dengan kebutuhan analisis sentimen calon bakal presiden 2024.

C . *Machine Translation*

Pada proses machine translation dilakukan untuk menerjemahkan kata berbahasa Indonesia ke dalam bahasa Inggris. Proses ini dilakukan apabila pelabelan data menggunakan library Vader Sentiment Lexicon. Oleh karena itu vader lexicon hanya mengandung kata berbahasa Inggris. Data yang akan diterjemahkan dengan machine translation menggunakan library googletrans.

```
import googletrans
from googletrans import Translator

translator = Translator()
translations = {}
for column in data.columns:
  unique_elements = data[column].unique()
  for element in unique_elements:
    translations[element] = translator.translate(element).text
translations
```

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

Figure 6. *Source code machine translation*

Tweet Cleaning	Hasil Machine Translation
berani jujur hebat capres melaksanakan keadilan sosial rakyat indonesia ganjar anies yg berpotensi tantangan politik global ganjar like anies	Dare to be honest, great presidential candidate implement social justice for the people of Indonesia, reward Anies who has the potential to challenge global politics, reward like Anies
safari politik dinilai etis ganjar baiknya pertimbangkan mundur gubernur	Political safaris are considered ethical as rewards for considering the governor's resignation
ganjar sesuaikan kerja jokowi kalo prabowo sorry susah umur	the reward is to adjust Jokowi's work if Prabowo is sorry it's hard to age

Table 5. *Sampel tweet hasil machine translation*

D . *Labelling Data*

Setelah melalui tahap pre-processing data selanjutnya pada tahap ini merupakan pelabelan data dengan menggunakan library VADER. VADER (Valanced Aware Dictionary Sentiment Reasoner) yang digunakan untuk secara otomatis melabeli data. Vader adalah pendekatakan leksikal yang digunakan sebagai model untuk analisis opini atau intensitas emosi dapat digunakan untuk menilai berbagai data. Kamus leksikon biasanya digunakan untuk mengevaluasi frasa dan kalimat sebagai sentimen tanpa perlu sumber lain berkonsultasi. Pada penelitian ini label data dalam bentuk multilabel yaitu label positif, negatif, dan netral. Pada proses labelling data library vader setiap tweet akan diberikan skor polaritas yang akan menunjukkan apakah termasuk klasifikasi label positif, negatif, atau netral.

	Tweet_Stopwords	Tweet_Clean	Positive	Negative	Neutral	Compound	Sentiment
0	dare to be honest, great presidential candidat...	dare honest, great presidential candidate impl...	0.647	0.000	0.353	0.9682	Positive
1	political safaris are considered ethical as re...	political safaris considered ethical rewards c...	0.471	0.162	0.368	0.6369	Positive
2	the reward is to adjust jokowi's work if prabo...	reward adjust jokowi's work prabowo sorry hard...	0.325	0.237	0.439	0.4588	Positive
3	try polling the empty box gp if your opponent ...	try polling empty box gp opponent rewards empt...	0.423	0.161	0.417	0.8957	Positive
4	anies' poll won, saying that it was only natur...	anies' poll won, saying natural anies won. vot...	0.557	0.094	0.348	0.9231	Positive
5	i love u neng agree reward president	love u neng agree reward president	0.839	0.000	0.161	0.8860	Positive

Figure 7. *Pemrosesan data tekstual dengan Vader Sentimen*

Jika nilai compound vader ini adalah hasil gabungan atau hasil dari nilai rata-rata bobot sentimen. Jika nilai compound >= 0.05 maka data ulasan tersebut merupakan sentimen positif. Jika nilai compound = 0 maka data ulasan tersebut termasuk sentimen netral. Apabila nilai compound <= 0.05 maka termasuk sentimen negatif.

```
score = data["Compound"].values
sentiment = []
for i in score:
    if i >= 0.05 :
        sentiment.append('Positive')
    elif i <= -0.05 :
        sentiment.append('Negative')
    else:
        sentiment.append('Neutral')
data["Sentiment"] = sentiment
data.head(17)
```

Figure 8. *Source code pemrosesan data tekstual*

Hasil dari pelabelan data setelah dilihat persentase dari jumlah data dengan kata kunci #GanjarPranowo sebanyak 975 terdapat 757 data yang memiliki sentimen positif, 114 data sentimen netral, dan 88 data sentimen negatif. Sedangkan jumlah data dengan kata kunci #AniesBaswedan sebanyak 367 terdapat 190 data yang memiliki sentimen positif, 89 data sentimen netral, dan 85 data sentimen negatif. Selain itu, jumlah data dengan kata kunci #PrabowoSubianto sebanyak 157 terdapat 95 data yang memiliki sentimen positif, 43 data sentimen netral, dan 14 data sentimen negatif.

E . Ekstraksi Fitur

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

Data hasil preprocessing berupa kata akan diubah menjadi angka dengan prosedur pembobotan kata kata yang bertujuan untuk menghitung bobot setiap kata yang akan dijadikan sebagai fitur, semakin banyak dokumen yang diproses maka maka semakin banyak fiturnya. Term weighting adalah proses pembobotan setiap kata untuk mengoptimalkan analisis sentimen dalam text mining. Penelitian ini menggunakan Term Frequency-Inverse Document Frequency (RF-IDF). Term Frequency (tf(w,d)) dianggap memiliki jumlah kemunculannya dalam teks atau dokumen. Inverse Document Frequency (IDF) adalah metode pembobotan token yang berfungsi untuk memonitor kemunculan token dalam sekumpulan teks. Pada Gambar 7 menunjukkan hasil perhitungan term frequency dan hasil perhitungan TF-IDF pada index ke 0.

```
vectorizer = TfidfVectorizer()
train_vectors = vectorizer.fit_transform(data_train['Tweet_Clean'])
test_vectors = vectorizer.transform(data_test['Tweet_Clean'])
```

Figure 9. *Pembobotan kata menggunakan TF-IDF*

```
print (train_vectors)

  (0, 2337)     0.2768099045469599
  (0, 1817)     0.1945123969596351
  (0, 2095)     0.42295222400907856
  (0, 1823)     0.1504626384400721
  (0, 948)      0.1274740613928334
  (0, 358)      0.19584215915493694
  (0, 1839)     0.1736321595750404
  (0, 1723)     0.32664855129295534
  (0, 1947)     0.3329529912087317
  (0, 1244)     0.19584215915493694
  (0, 2655)     0.35721474046130186
  (0, 1005)     0.2500015226183187
  (0, 1283)     0.2768099045469599
  (0, 2048)     0.2768099045469599
  (1, 2322)     0.3838349422958192
  (1, 1964)     0.4992619508468104
  (1, 1673)     0.5296438361180652
  (1, 2593)     0.3893879229657591
  (1, 2039)     0.26009381516553315
  (1, 2672)     0.3218911324230471
  (2, 876)      0.4121349187478304
```

Figure 10. *Hasil perhitungan TF-IDF*

F . Klasifikasi Support Vector Machine (SVM)

Support Vector Machine (SVM) salah satu metode klasifikasi dengan menggunakan machine learning (supervised learning) yang memprediksi kelas berdasarkan model atau sampel hasil dari proses pelatihan atau training. Klasifikasi dilakukan dengan mencari hyperplane atau pembatas (decision boundary) memisahkan kelas dengan kelas lain, dalam hal ini garis tersebut memisahkan tweet sentimen positif dengan sentimen negatif atau netral. Akan tetapi sebelum melakukan klasifikasi data dibagi menjadi data training dan data testing. Proses pembagian data menggunakan modul train_test_split dari library sklearn.model_selection dengan python. Rasio pembagian data latih dan data uji adalah 90:10, dimana 90% untuk data latih dan 10% untuk data uji. Data latih digunakan untuk membangun model atau mendefinisikan pola. Setelah data dibagi menjadi dua bagian lalu dapat digunakan pada proses klasifikasi menggunakan algoritma Support Vector Machine (SVM). Pada tabel 7 menunjukkan bahwa hasil dari klasifikasi menggunakan dataset kata kunci #ganjarpranowo tingkat akurasi metode SVM yaitu 84%.

	Accuracy	Precision	Recall
SVM Kernel = Linear	0,84375	0,689	0,569
SVM Kernel = RBF	0,8229	0,6057	0,424
SVM Kernel = Polynomial	0,8028	0,6	0,363
SVM Kernel = Sigmoid	0,802	0,491	0,389

Figure 11. *Hasil klasifikasi SVM dataset Ganjar Pranowo*

Pada tabel 8 menunjukkan bahwa hasil dari klasifikasi menggunakan dataset kata kunci #prabowosugianto tingkat akurasi metode SVM yaitu 56%.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol 8 No 1 (2023): June
DOI: 10.21070/acopen.8.2023.7138 . Article type: (Political science)

	Accuracy	Precision	Recall
SVM Kernel = Linear	0,5625	0,375	0,399
SVM Kernel = RBF	0,75	0,5714	0,466
SVM Kernel = Polynomial	0,625	0,2083	0,333
SVM Kernel = Sigmoid	0,75	0,5714	0,466

Figure 12. *Hasil klasifikasis SVM dataset Prabowo Sugianto*

Pada tabel 9 menunjukkan bahwa hasil dari klasifikasi menggunakan dataset kata kunci #aniesbaswedan tingkat akurasi metode SVM yaitu 62%.

	Accuracy	Precision	Recall
SVM Kernel = Linear	0,6216	0,65	0,59
SVM Kernel = RBF	0,5945	0,4157	0,4639
SVM Kernel = Polynomial	0,5135	0,1717	0,3333
SVM Kernel = Sigmoid	0,5405	0,5092	0,2969

Figure 13. *Hasil klasifikasiS SVM dataset Anies Baswedan*

G . Klasifikasi Bernoulli Naïve Bayes (BNB)

Algortima Bernoulli Naïve Bayes digunakan mengimplementasikan klasifikasi data yang terdistribusi Bernolli multivariat yaitu mungkin terdapat beberapa fitur sekalipun masing-masing diperlakukan sebagai variabel nilai biner (Bernoulli, boolean). Hasil dari algoritma klasifikasi BNB menyatakan bahwa tingkat akurasi pada metode ini dapat dilihat pada Tabel 10 menunjukkan bahwa data dengan kata kunci #ganjarpranowo mendapatkan nilai akurasi sebesar 77%.

	Accuracy	Precision	Recall
Bernoulli Naïve Bayes	0,7708	0,2652	0,325

Figure 14. *Hasil klasifikasi BNB dataset Ganjar Pranowo*

Pada tabel 11 menunjukkan bahwa hasil dari klasifikasi menggunakan dataset kata kunci #prabowosugianto tingkat akurasi metode BNB yaitu 43%.

	Accuracy	Precision	Recall
Bernoulli Naïve Bayes	0,4375	0,2555	0,266

Figure 15. *Hasil klasifikasi BNB Dataset Prabowo Sugianto*

Pada tabel 12 menunjukkan bahwa hasil dari klasifikasi menggunakan dataset kata kunci #prabowosugianto tingkat akurasi metode BNB yaitu 45%.

	Accuracy	Precision	Recall
Bernoulli Naïve Bayes	0,4594	0,3387	0,3177

Figure 16. *Hasil klasifikasi BNB dataset Anies Baswedan*

H. Klasifikasi Logistic Regression (LR)

	Accuracy	Precision	Recall
Logistic Regression	0,84375	0,7835	0,55

Figure 17. *Dataset Ganjar Pranowo*

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

	Accuracy	Precision	Recall
Logistic Regression	0,5625	0,344	0,366

Table 6. *Dataset Prabowo Sugianto*

	Accuracy	Precision	Recall
Logistic Regression	0,4594	0,3387	0,3177

Figure 18. *Dataset Anies Baswedan*

Berdasarkan tabel di atas pada 3 dataset yaitu dataset Ganjar pranowo, Dataset Anies baswedan maupun dataset Prabowo Subianto, akurasi tertinggi pada saat dilakukan klasifikasi yaitu menggunakan metode Logistic regression dan SVM menggunakan kernel linier pada dataset Ganjar Pranowo. Hal ini menunjukkan bahwa pada dataset Prabowo Subianto maupun dataset Anies Baswedan memiliki leksikon kata yang tidak teridentifikasi pada sentimen negatif. Hal ini disebabkan pada saat dilakukan labelling menggunakan Vader sentiment terdapat kosakata yang tidak ditampung di dalam vader sentimen, sehingga yang seharusnya menjadi kata negatif namun diartikan positif. Hal ini ditemukan pada dataset Anies Basewdan yaitu kata “Nyinyir” jika diterjemahkan ke bahasa inggris yaitu “smirk” , pada bahasa indonesia merupakan konotasi negatif, namun pada bahasa inggris tidak dimasukkan ke dalam leksikon kata negatif pada daftar vader sentimen. Selain itu terdapat kata “Anjing” jika diterjemahkan ke bahasa inggris yaitu dog , Kata tersebut tidak termasuk di daftar ke dalam leksikon kata negatif karena dog bukan kata hinaan di dalam bahasa inggris tapi lebih merujuk ke hewan. Selain itu terdapat beberapa leksikon kata negatif yang tidak ditampung di vader sentimen yaitu “ Kadrun”, “Tukang ngibul” .

Simpulan

Dari hasil menunjukan bahwa pengelompokan sentimen positif dan negatif pada anotasi otomatis menggunakan vader sentimen , terdapat kata-kata bahasa indonesia yang diterjermahkan menjadi bahasa inggris yang seharusnya masuk sentimen negatif namun masuk ke sentimen positif. Sehingga hal tersebut berpengaruh terhadap hasil klasifikasi , hal ini ditunjukkan bahwa dataset prabowo maupun anies hasil akurasinya cukup rendah karena banyak kata-kata leksikon yang tidak ditampung pada vader sentiment. Selaint itu kata-kata negatif cukup berbeda pada tiap negara sehingga berpengaruh pada hasil akurasi. Saran pada penelitian berikutnya anotasi label otomatis bisa menggunakan metode machine learning semisupervised dan klasifikasi mengunakan metode deep learning serta menambahakan leksikon-leksikon kata negatif pada daftar kata sentimen negatif

References

1. D. A. Vonega, A. Fadila, and D. E. Kurniawan, “Analisis Sentimen Twitter Terhadap Opini Publik Atas Isu Pencalonan Puan Maharani dalam PILPRES 2024,” J. Appl. Informatics Comput., vol. 6, no. 2, pp. 129–135, 2022, doi: 10.30871/jaic.v6i2.4300.
2. I. Kurniawan and A. Susanto, “Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019,” Eksplora Inform., vol. 9, no. 1, pp. 1–10, 2019, doi: 10.30864/eksplora.v9i1.237.
3. G. Sanjaya and K. M. Lhaksmana, “Analisis Sentimen Komentar YouTube tentang Terpilihnya Menteri Kabinet Indonesia Maju Menggunakan Lexicon Based,” vol. 7, no. 3, pp. 9698–9710, 2020.
4. A. D. Akmal, I. Permana, H. Fajri, and Y. Yuliarti, “Opini Masyarakat Twitter terhadap Kandidat Bakal Calon Presiden Republik Indonesia Tahun 2024,” J. Manaj. dan Ilmu Adm. Publik, vol. 4, no. 4, pp. 292–300, 2022, doi: 10.24036/jmiap.v4i4.160.
5. Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, “Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile,” Petir, vol. 15, no. 2, pp. 264–275, 2022, doi: 10.33322/petir.v15i2.1733.
6. W. Athira Luqyana, I. Cholissodin, and R. S. Perdana, “Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine,” J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 11, pp. 4704–4713, 2018, [Online]. Available: http://j-ptiik.ub.ac.id
7. M. Z. Anbari and B. Sugiantoro, “Studi Komparasi Metode Analisis Sentimen Naïve Bayes , SVM , dan Logistic Regression Pada Piala Dunia 2022,” vol. 7, no. April, pp. 688–695, 2023, doi: 10.30865/mib.v7i2.5383.
8. N. S. Wardani, A. Prahutama, and P. Kartikasari, “Analisis Sentimen Pemindahan Ibu Kota Negara Dengan Klasifikasi Naïve Bayes Untuk Model Bernoulli Dan Multinomial,” J. Gaussian, vol. 9, no. 3, pp. 237–246, 2020, doi: 10.14710/j.gauss.v9i3.27963.
9. Primi Rohimi, “SNA DENGAN NETLYTIC PADA KOLOM KOMENTAR VIDEO YOUTUBE GUS MIFTAH CERAMAH DI GEREJA,” vol. 1, no. 1, pp. 360–377, 2021, doi: 10.21154/dialogia.v15i2.1192.4.
10. B. M. Pintoko and K. M. L, “Analisis Sentimen Jasa Transportasi Online pada Twitter Menggunakan Metode

Naïve Bayes Classifier," 2018.
11. N. Putu, A. Widiari, I. M. Agus, D. Suarjaya, and D. P. Githa, "Teknik Data Cleaning Menggunakan Snowflake untuk Studi Kasus Objek Pariwisata di Bali," vol. 8, no. 2, pp. 137–145, 2020.
12. V. A. Flores and L. Jasa, "Analisis Sentimen untuk Mengetahui Kelemahan dan Kelebihan Pesaing Bisnis Rumah Makan Berdasarkan Komentar Positif dan Negatif di Instagram," vol. 19, no. 1, 2020.
13. D. W. Seno and A. Wibowo, "Analisis Sentimen Data Twitter Tentang Pasangan Capres-Cawapres Pemilu 2019 Berbasis Metode Lexicon Dan Support Vector Machine," vol. XI, no. 2, pp. 144–155, 2019.
14. I. T. Julianto, "Analisis Sentimen Terhadap Sistem Informasi Akademik Mahasiswa Institut Teknologi Garut," pp. 458–465.
15. A. Novantirani et al., "Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine," pp. 1–7, 2015.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo